MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Approaches for Low-resource Neural Machine Translation

Feng Hou

Ruili Wang

Content


Data Augmentation Approaches
Model Training Approaches
Our Work
Word-level
Meta-learning
Sentence-level
Multilingual
Diversified
Unsupervised

Data Augmentation Approaches

1. Word-level Data Augmentation

- Replace with rare words (Fadaee et al., 2017)
- Replace random words with other words in the vocabularies SwitchOut (Wang et al., 2018)
- Soft augmentation (Gao et al., 2019)



A boy is holding a bat.

computer vision augmentation



A boy is holding a bat.

A boy is holding a bat.
Ein Junge hält einen Schläger.

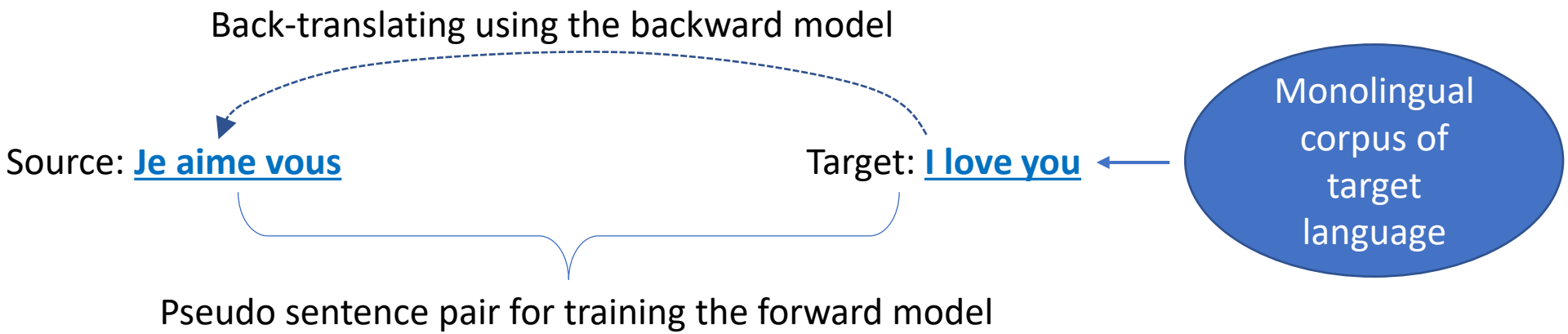
translation augmentation

A boy is holding a **backpack**.
Ein Junge hält einen *Rucksack*.

Data Augmentation Approaches

2. Sentence-level Data Augmentation

- Back-translation (Sennrich et al., 2016a)
- Iterative back-translation (Hoang et al., 2018)
- Meta back-translation (Pham et al., 2021)



Data Augmentation Approaches

3. Diversified Data Augmentation

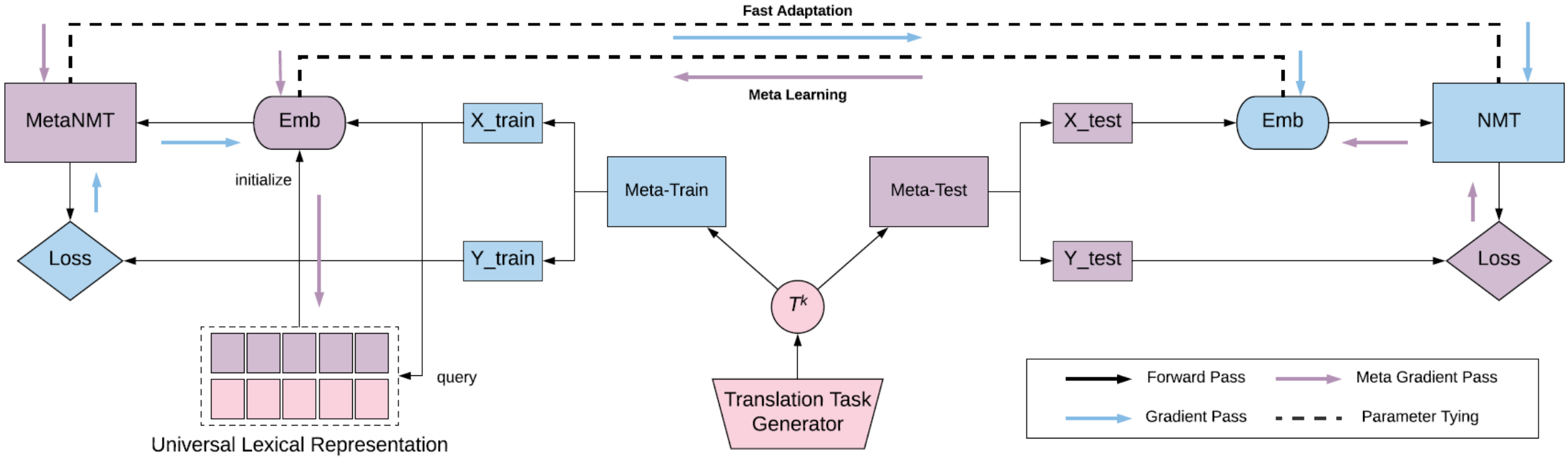
- Combinations of back-translated data from PB-SMT and NMT (Poncelas et al., 2019)
- Data diversification (Nguyen et al., 2020)
- Selecting back-translated data from multiple sources: rule-based, PB-SMT and NMT (Soto et al., 2020)
- Dynamic data selection and weighting for iterative back-translation (Dou et al., 2020)

lower-quality but diverse data often yields stronger results

Model Training Approaches

1. Meta-learning Approach

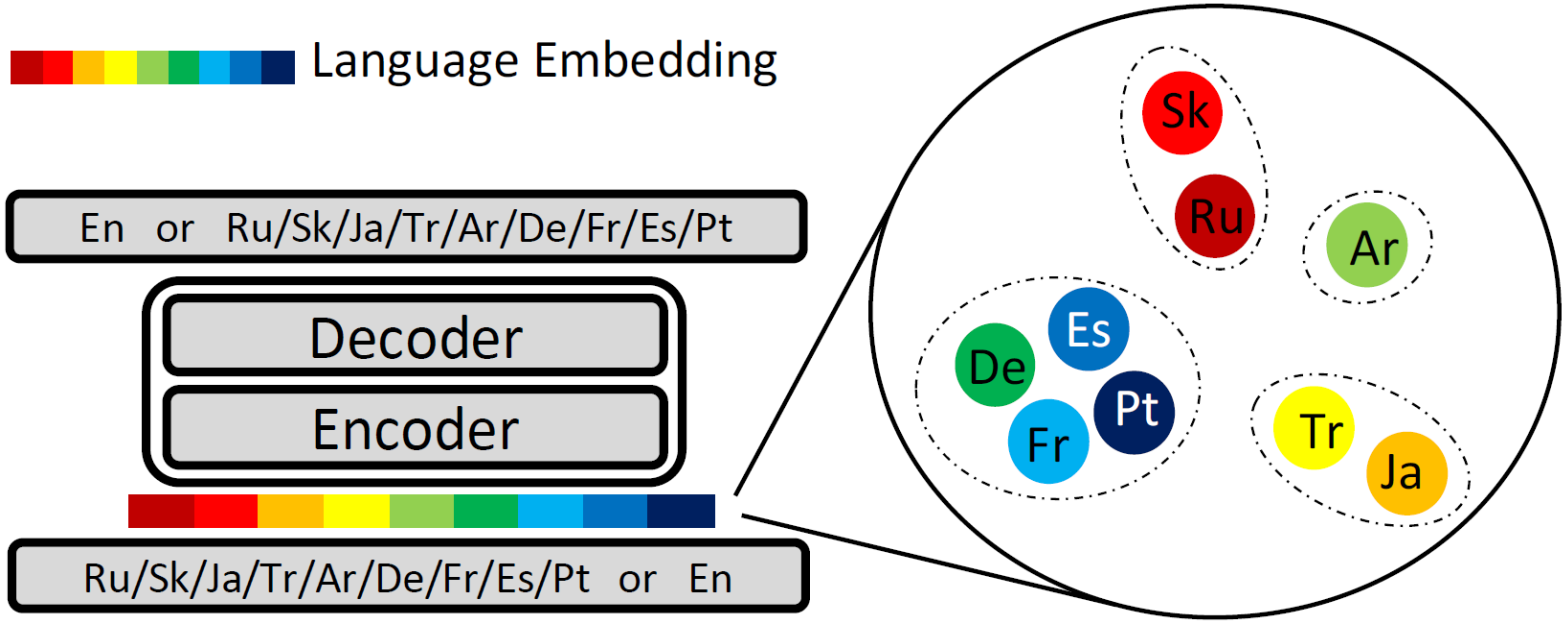
- Using auxiliary translation tasks (Gu et al.,2018c)
- **Cannot use monolingual data (only translation tasks are considered)**



Model Training Approaches

2. Multilingual Approach

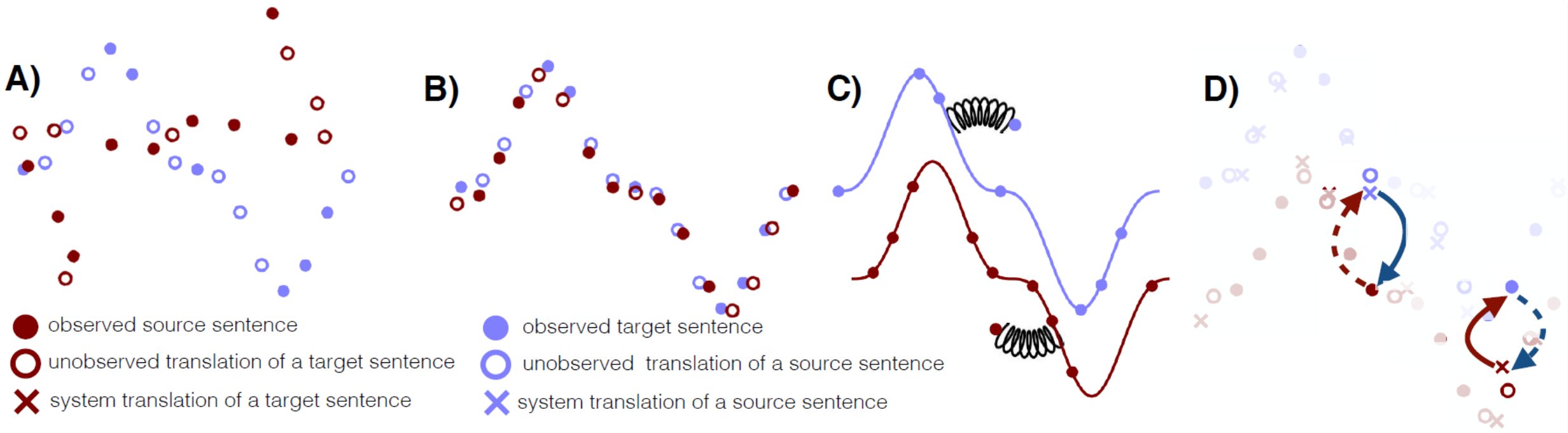
- Decoder uses language embeddings (Tan et al., 2019)
- Incorporate multi-task learning (Wang et al., 2020b), de-noising auto-encoder pre-training (Liu et al., 2020a) or transfer learning (Zoph et al., 2016).
- **Needs larger model structures**



Model Training Approaches

3. Unsupervised Approach

- Shared encoder representations (Lample et al., 2018)
- Pre-trained cross-lingual language model (Lample and Conneau, 2019).
- **Impaired when languages are morphologically different**



Our Work

Tested ideas on Ro-En

- Diversified data augmentation: incorporating paraphrase generation **(Mediocre)**
- Metric learning of the encoder **(Negative)**
- Function words dropout: using IDF to identify function words **(Minor)**

Working on

- Refine data augmentation
- Supervised label smoothing: unsupervised label smoothing assign small probability mass to randomly selected other words.
- Monolingual Meta-learning: incorporating training tasks(e.g., MLM, Auto-encoder and paraphrase generation other than translation tasks) of the same language